ENDINGS: USING DIAGNOSTICS

Joey Takeda

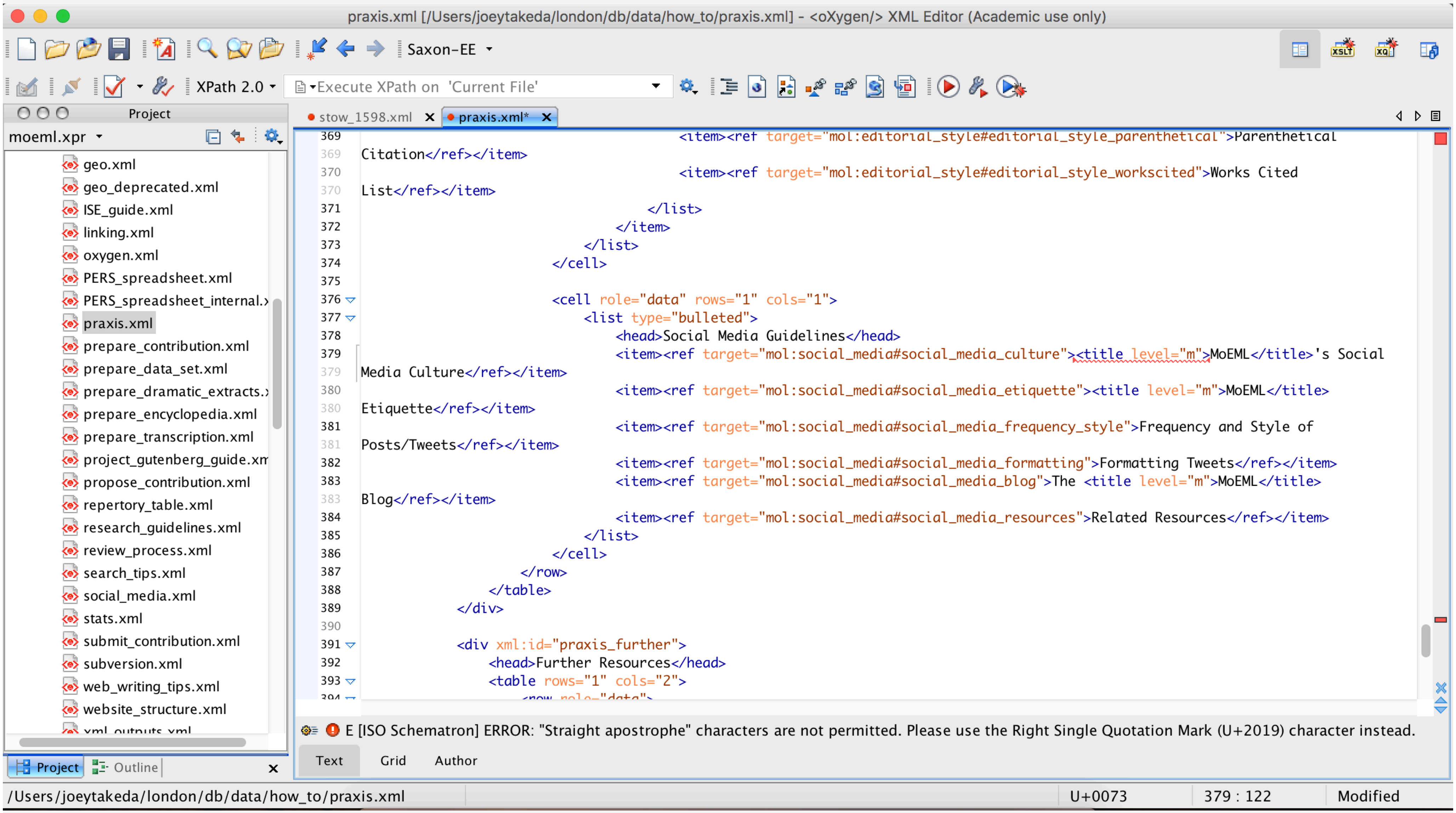
Junior Programmer

WHAT ARE DIAGNOSTICS

- A generated file (usually HTML or text) that displays common encoding errors and problems that are not usually caught by schema/schematron
- Validation catches for local errors in the data encoded in an original file and does not usually analyze linked data
- Diagnostics interrogate the linked data between documents, while also checking local errors, issues of consistency, and potential errors

WHAT VALIDATION CAN DO

- Check that all ids are in the proper form (eg: if the file is named “BLOG2” then any ids created further in the file must start with “BLOG2”)
- Ensure that elements of the file are included before particular stages of the document (eg: All published location files **must** include an abstract)
- Enforce stylistic conformity (eg: Straight apostrophes are not allowed, use curly apostrophes)
- +++++++



praxis.xml [/Users/joeytakeda/london/db/data/how_to/praxis.xml] - <oXygen/> XML Editor (Academic use only)

- geo.xml
- geo_deprecated.xml
- ISE_guide.xml
- linking.xml
- oxygen.xml
- PERS_spreadsheet.xml
- PERS_spreadsheet_internal.x
- praxis.xml
- prepare_contribution.xml
- prepare_data_set.xml
- prepare_dramatic_extracts.x
- prepare_encyclopedia.xml
- prepare_transcription.xml
- project_gutenberg_guide.xm
- propose_contribution.xml
- repertory_table.xml
- research_guidelines.xml
- review_process.xml
- search_tips.xml
- social_media.xml
- stats.xml
- submit_contribution.xml
- subversion.xml
- web_writing_tips.xml
- website_structure.xml
- xml_outputs.xml

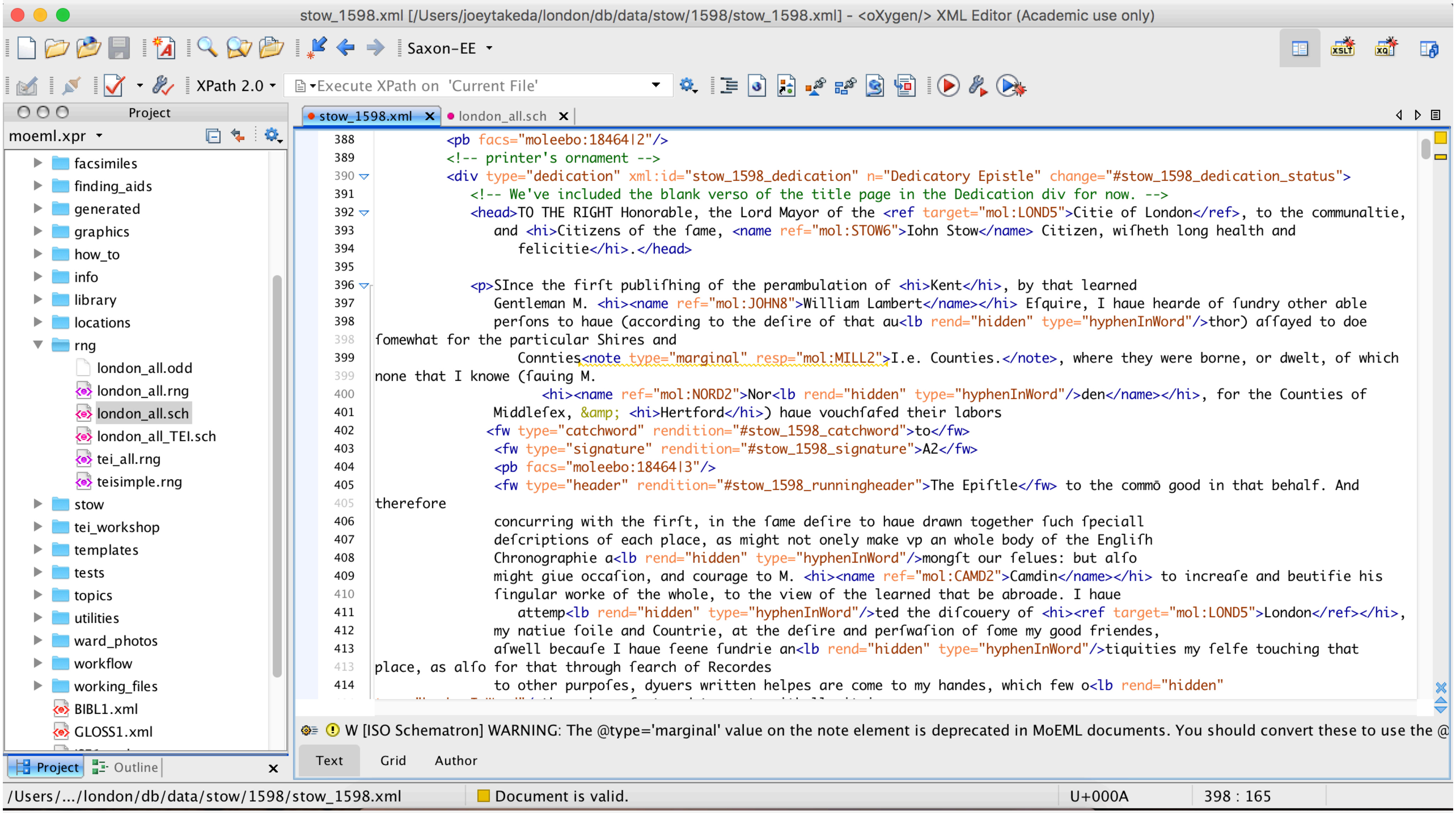
```
                    <item><ref target="mol:editorial_style#editorial_style_parenthetical">Parenthetical
Citation</ref></item>
                    <item><ref target="mol:editorial_style#editorial_style_workscited">Works Cited
List</ref></item>
                  </list>
                </item>
              </list>
            </cell>

            <cell role="data" rows="1" cols="1">
              <list type="bulleted">
                <head>Social Media Guidelines</head>
                <item><ref target="mol:social_media#social_media_culture"><title level="m">MoEML</title>'s Social
Media Culture</ref></item>
                <item><ref target="mol:social_media#social_media_etiquette"><title level="m">MoEML</title>
Etiquette</ref></item>
                <item><ref target="mol:social_media#social_media_frequency_style">Frequency and Style of
Posts/Tweets</ref></item>
                <item><ref target="mol:social_media#social_media_formatting">Formatting Tweets</ref></item>
                <item><ref target="mol:social_media#social_media_blog">The <title level="m">MoEML</title>
Blog</ref></item>
                <item><ref target="mol:social_media#social_media_resources">Related Resources</ref></item>
              </list>
            </cell>
          </row>
        </table>
      </div>

      <div xml:id="praxis_further">
        <head>Further Resources</head>
        <table rows="1" cols="2">
          <row role="data">
```

E [ISO Schematron] ERROR: "Straight apostrophe" characters are not permitted. Please use the Right Single Quotation Mark (U+2019) character instead.

/Users/joeytakeda/london/db/data/how_to/praxis.xml U+0073 379 : 122 Modified

WHAT IT CAN’T (OR SHOULDN’T)

- Check to see if an @xml:id is unique out of 10742 other ids across over 1600 files
- Test to see if an a particular pointer actually points to something in another file
- Double check if the thing you’re referencing is actually what you’re claiming it is (eg:tagging a person as a location)
- Give a suggestion;even though you can differentiate between warnings and errors in schematron,the ominous line still makes you think what you’re doing might be wrong even if it isn’t



stow_1598.xml [/Users/joeytakeda/london/db/data/stow/1598/stow_1598.xml] - <oXygen/> XML Editor (Academic use only)

```
        <pb facs="moleebo:18464|2"/>
        <!-- printer's ornament -->
        <div type="dedication" xml:id="stow_1598_dedication" n="Dedicatory Epistle" change="#stow_1598_dedication_status">
           <!-- We've included the blank verso of the title page in the Dedication div for now. -->
           <head>TO THE RIGHT Honorable, the Lord Mayor of the <ref target="mol:LOND5">Citie of London</ref>, to the communaltie,
              and <hi>Citizens of the ſame, <name ref="mol:STOW6">Iohn Stow</name> Citizen, wiſheth long health and
                 felicitie</hi>.</head>

           <p>SInce the firſt publiſhing of the perambulation of <hi>Kent</hi>, by that learned
              Gentleman M. <hi><name ref="mol:JOHN8">William Lambert</name></hi> Eſquire, I haue hearde of ſundry other able
              perſons to haue (according to the deſire of that au<lb rend="hidden" type="hyphenInWord"/>thor) aſſayed to doe
ſomewhat for the particular Shires and
                 Conntieſ<note type="marginal" resp="mol:MILL2">I.e. Counties.</note>, where they were borne, or dwelt, of which
none that I knowe (ſauing M.
                    <hi><name ref="mol:NORD2">Nor<lb rend="hidden" type="hyphenInWord"/>den</name></hi>, for the Counties of
              Middleſex, &amp; <hi>Hertford</hi>) haue vouchſafed their labors
             <fw type="catchword" rendition="#stow_1598_catchword">to</fw>
              <fw type="signature" rendition="#stow_1598_signature">A2</fw>
              <pb facs="moleebo:18464|3"/>
              <fw type="header" rendition="#stow_1598_runningheader">The Epiſtle</fw> to the commō good in that behalf. And
therefore
              concurring with the firſt, in the ſame deſire to haue drawn together ſuch ſpeciall
              deſcriptions of each place, as might not onely make vp an whole body of the Engliſh
              Chronographie a<lb rend="hidden" type="hyphenInWord"/>mongſt our ſelues: but alſo
              might giue occaſion, and courage to M. <hi><name ref="mol:CAMD2">Camdin</name></hi> to increaſe and beutifie his
              ſingular worke of the whole, to the view of the learned that be abroade. I haue
                 attemp<lb rend="hidden" type="hyphenInWord"/>ted the diſcouery of <hi><ref target="mol:LOND5">London</ref></hi>,
              my natiue ſoile and Countrie, at the deſire and perſwaſion of ſome my good friendes,
              aſwell becauſe I haue ſeene ſundrie an<lb rend="hidden" type="hyphenInWord"/>tiquities my ſelfe touching that
place, as alſo for that through ſearch of Recordes
              to other purpoſes, dyuers written helpes are come to my handes, which few o<lb rend="hidden"
```

W [ISO Schematron] WARNING: The @type='marginal' value on the note element is deprecated in MoEML documents. You should convert these to use the @

/Users/.../london/db/data/stow/1598/stow_1598.xml

Document is valid.

U+000A

398 : 165

MOEML’S DIAGNOSTICS

- Diagnostics are generated as part of the regular build process
- Checks the encoding of the data, especially in relation to the data points to which the encoding links
- Checks for possible or “fuzzy” errors in the data: things that a computer can flag up as potentially wrong, but requires human intervention

To Jenkins….

ENCODING ERRORS

- Duplicate ids (breaks the build completely)
- <name> instead of <ref>
- Organizations tagged as <name>s but without an @type='org'
- Bibliography entries tagged without @type='bibl'
- Malformed ids (mostly typos)

BENEFITS

- While validation scenarios could potentially check for some errors like duplicate ids or malformed ids, it would require significant processing energy
- Catches typos and forces editors to go back and review their work, thus operates as a form of proofing
- Gives a graphic representation on how "healthy" the data is
- An easy way for large projects to chip away at legacy code
  - Every team member takes 10 minutes out at the end of the day to fix a few errors

“FUZZY” ERRORS

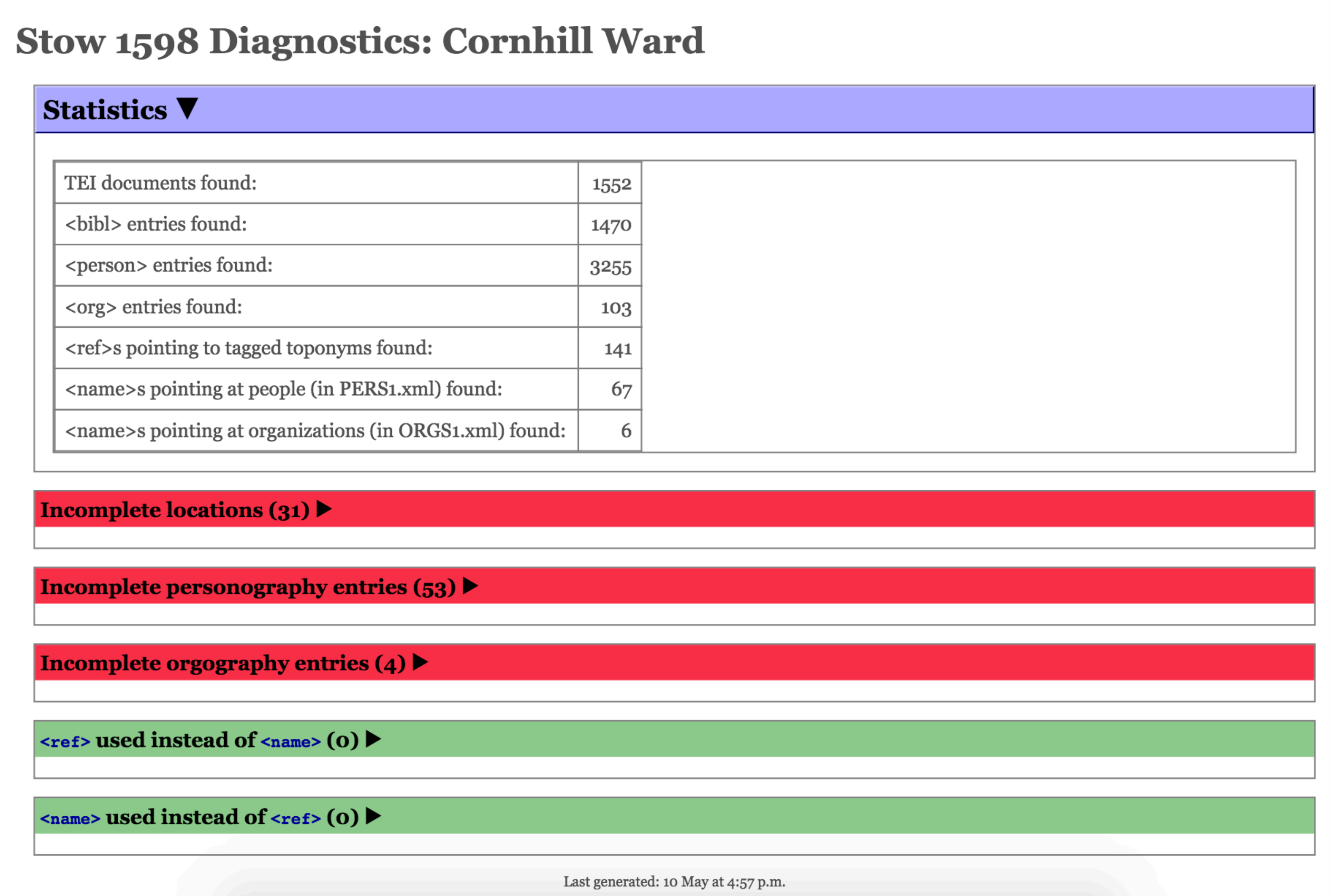
- These are issues that aren’t technically problems, but would make the site better if someone would fix or address them

BENEFITS

- Checks for possible errors without making something that is correct invalid
- Allows for decisions about project practices to be made and implemented as time allows, especially for site-wide changes that would cause the majority of the project to become invalid
- **Problem**: How do we declare in our encoding that something that might look incorrect is actually correct?

CASE STUDY: STOW_1598

- Alongside our implementation of the regular site diagnostics, we have also created a similar system for our process of proofing the 1598 edition of Stow
- Proofing a text of this size is a challenge, especially since it requires that we have good definitions of what it means for something to be "finished" on the site
- Before we send chapters of Stow to peer-review (declaring that it is "finished"), we must ensure that everything that is included is also finished



Stow 1598 Diagnostics: Cornhill Ward

Statistics ▼

TEI documents found:	1552
<bibl> entries found:	1470
<person> entries found:	3255
<org> entries found:	103
<ref>s pointing to tagged toponyms found:	141
<name>s pointing at people (in PERS1.xml) found:	67
<name>s pointing at organizations (in ORGS1.xml) found:	6

Incomplete locations (31) ▶

Incomplete personography entries (53) ▶

Incomplete orgography entries (4) ▶

`<ref>` used instead of `<name>` (0) ▶

`<name>` used instead of `<ref>` (0) ▶

Last generated: 10 May at 4:57 p.m.

http://jenkins.hcmc.uvic.ca/job/MoEML/lastSuccessfulBuild/artifact/products/stow_1598_CORN1_diagnostics.htr

CONCLUSION

- Diagnostics:
  - Allow encoders to check their work and make sure everything will function the way they want it to
  - Maintain site consistency and usability
  - Give a rendering of data's readiness for proofing/publishing purposes
- However, diagnostics can be misleading if there is no standardized way of declaring that an item has been checked and is certainly not an error
- Going forward: milestones? Graphic representations? Timelines?